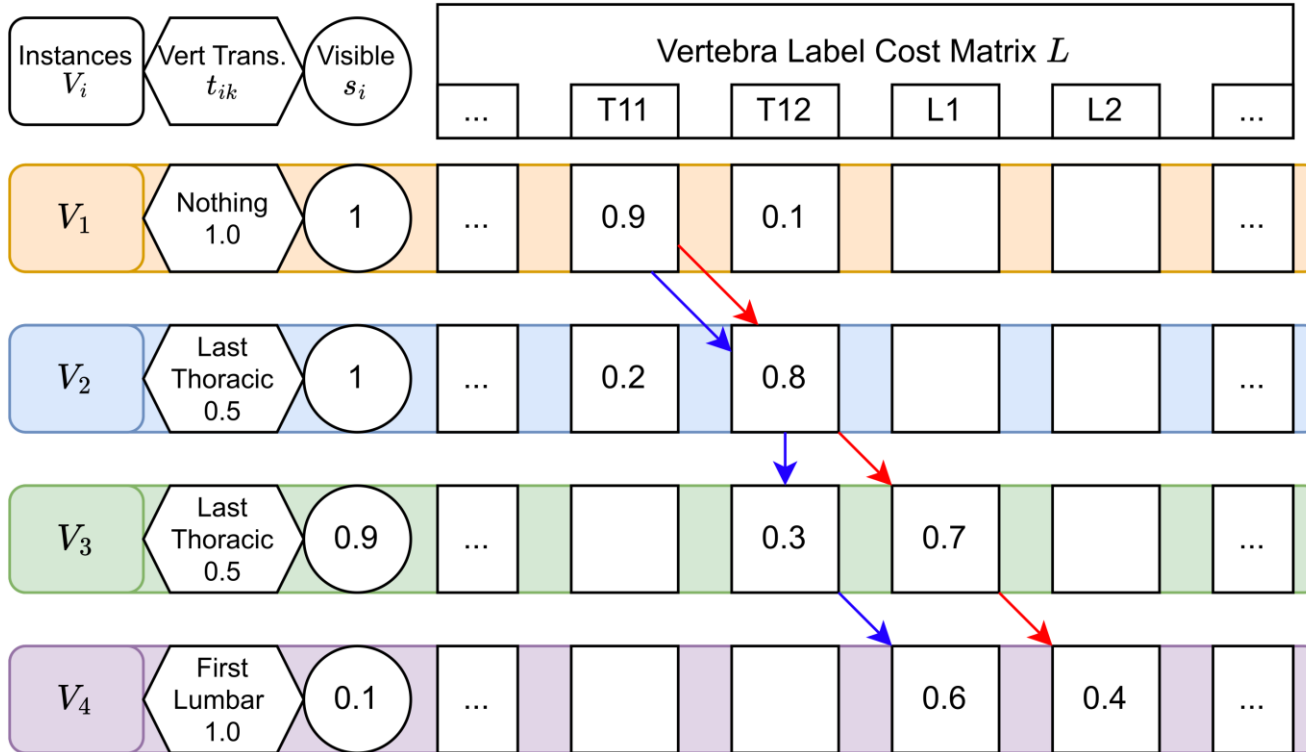


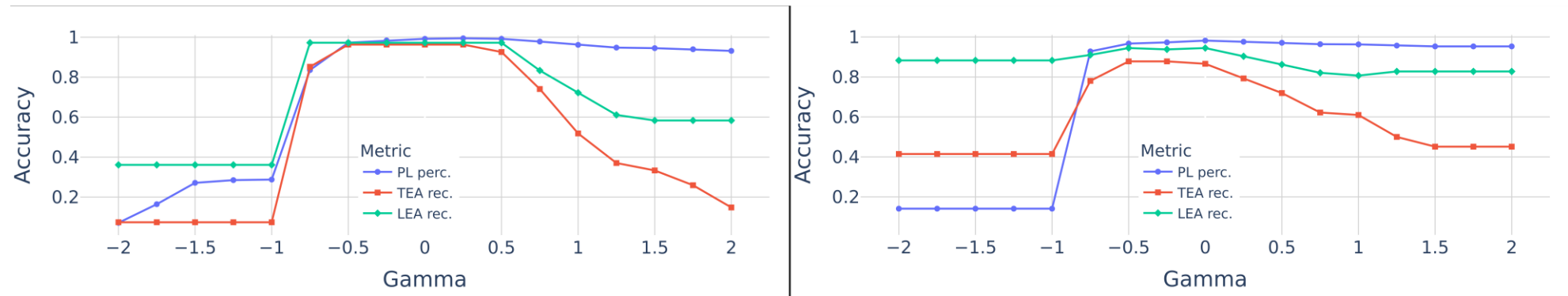
VERIDAH: Solving Enumeration Anomaly Aware Vertebra Labeling across Imaging Sequences

ELECTRONIC SUPPLEMENTARY MATERIAL

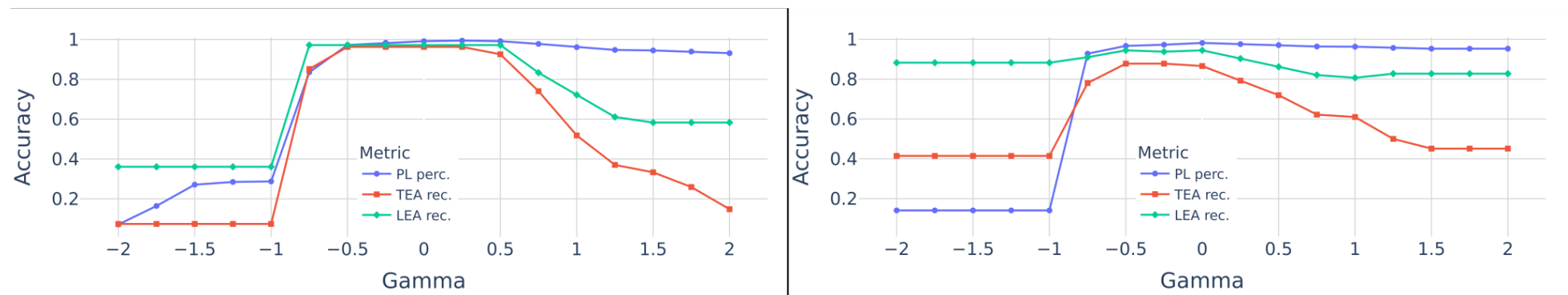
Supplemental Fig. 1: Example of our sequence predictor on an artificially created sample containing $n = 4$ vertebrae ($V_1 - V_4$) in the FOV. The classifier outputs specific vertebra class predictions c_{ij} and region prediction scores r_{ij} for each class (already combined into the label cost matrix L and reduced to T11 to L2 for simplicity), their visibility s_i , and the vertebra transitions output t_{ik} . Depending on the utilized classification head, either the path [T11, T12, L1, L2] (red arrows) or [T11, T12, T12, L1] (blue arrows) is returned by the sequence predictor. In the case of the latter one, the post-processing relabels the path [T11, T12, T12, L1] to [T11, T12, T13, L1], making this a TEA case. All unspecified classifier outputs have a value of zero.



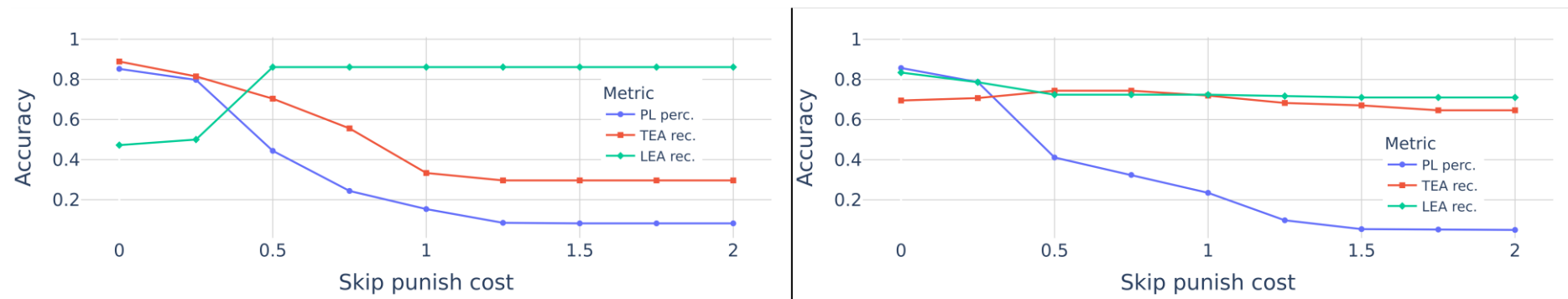
Supplemental Fig. 2: The behavior of three metrics in relation to a specified artificial cost punishment of γ on our CT (left) and MRI (right) test data. We sample these results in 0.25 steps, as indicated by the plot markers. The baseline is represented by $\gamma = 0$.



Supplemental Fig. 3: The behavior of three metrics, if we allow arbitrary label gaps in our sequence predictor on the CT (left) and MRI (right) test data. The x-axis shows different costs to punish these skips. We sample these results in 0.25 steps, as indicated by the plot markers.



Supplemental Fig. 4: The behavior of three metrics, if we allow arbitrary label gaps in our sequence predictor on the CT (left) and MRI (right) test data, and we remove exactly one random vertebra in each subject. The x-axis shows different costs to punish these skips. We sample these results in 0.25 steps, as indicated by the plot markers.



Supplemental Table 1. Training Hyperparameters. Different setups for the sequences are written with slashes (T2w sag / CT)

Parameter	Value
Data	
Number of 3D scans	5291 / 1171
Orientation	(Posterior, Inferior, Right)
Resolution (mm)	(0.85, 0.85, 3.3) / (1.0, 1.0, 1.0)
Crop Size (voxel count)	(200, 160, 32) / (128, 128, 128)
Training Setup	
Architecture	DenseNet 169
Batch Size	32
Max Epochs	100
LR	1e-4
LR End Factor	0.01
L2 Regularization	1e-3
Dropout	0.1
Data Augmentation	
RandCut	prob=0.20
RandAffine	prob=0.40, translate=8, scale=0.15, degrees=15°
RandScaleIntensity	prob=0.15, offset=0.15
RandShiftIntensity	prob=0.15, factor=0.15
RandGaussianSmooth	prob=0.10, sigma=0.60
RandGaussianNoise	prob=0.10, std=10.0
RandAdjustContrast	prob=0.10, gamma=1.0
RandLowResolution	prob=0.10, range=(0.8, 1.0)

Supplemental Table 2. Labeling Performance of the VERIDAH models on abnormal MRI (scoliosis or stump rib) and CT (fractures, stump rib) subsets of the test data. Abbreviations: rec: recall, prec: precision, TEA: Thoracolumbar Enumeration Anomaly, LEA: Lumbar Enumeration Anomaly.

Metric	Total test set	Scoliosis	Stump Rib
T2w sagittal			
Number of test subjects	1000	62	194
PL perc.	98.3	96.77	93.81
Sub. Corr.	99.58 ± 3.19	99.23 ± 4.22	98.49 ± 5.87
TEA rec.	87.8	100	84.44
LEA rec.	94.48	80	88.89
TEA prec.	89.87	71.43	88.37
LEA prec.	92.57	100	80
CT			
	Total test set	Fractures	Stump Rib
Number of test subjects	365	81	46
PL perc.	99.18	97.53	93.48
Sub. Corr.	99.63 ± 4.37	98.72 ± 8.52	97.07 ± 10.99
TEA rec.	96.3	85.71	96
LEA rec.	97.22	100	95
TEA prec.	92.86	85.71	92.31
LEA prec.	97.22	88.89	95

Supplemental Table 3. Labeling Performance of the VERIDAH models on the MRI test data, stratified by the four most common thoracolumbar enumeration counts. Enumeration is denoted thoracic enumeration and then lumbar separated by a slash (e.g. T12/L5 means twelve thoracic and five lumbar vertebrae). Abbreviations: rec: recall, prec: precision, TEA: Thoracolumbar Enumeration Anomaly, LEA: Lumbar Enumeration Anomaly.

Metric	T12/L5	T12/L6	T13/L5	T11/L6
Number of test subjects	796	107	46	21
PL perc.	99.75	98.13	80.43	90.48
Sub. Corr.	99.93 $\pm$ 1.31	99.53 $\pm$ 3.39	95.11 $\pm$ 9.92	97.52 $\pm$ 7.66
TEA rec.	N/A	N/A	80.43	90.48
LEA rec.	N/A	98.13	N/A	90.48
TEA prec.	N/A	N/A	100	100
LEA prec.	N/A	100	N/A	100

Supplemental Table 4. Labeling Performance T2w sagittal compared to training without labeled anomalies. Abbreviations: rec: recall, prec: precision, TEA:

Thoracolumbar Enumeration Anomaly, LEA: Lumbar Enumeration Anomaly. 4H-P-NA: Taking all four classification heads with the sequence predictor using no TEA or LEA labels during training. VH-P-NA: Using only the vertebra output and the sequence predictor using no TEA or LEA labels during training.

Metric	VERIDAH	4H-P-NA	VH-P-NA
		T2w sagittal	
PL perc.	98.3	93.4	93.1
Sub. Corr.	99.58 ± 3.19	98.32 ± 6.52	98.22 ± 6.96
TEA rec.	87.8	32.93	39.02
LEA rec.	94.48	82.76	77.24
TEA prec.	89.87	71.05	61.54
LEA prec.	92.57	75	75.68
		CT	
PL perc.	99.18	93.7	92.6
Sub. Corr.	99.63 ± 4.37	98.08 ± 7.77	97.78 ± 8.24
TEA rec.	96.3	25.93	22.22
LEA rec.	97.22	63.89	52.78
TEA prec.	92.86	70	50
LEA prec.	97.22	71.88	67.86

Supplemental Materials

Appendix A: VERIDAH Mathematical Details

Let $V = \{V_1, V_2, \dots, V_n\}$ be the sequence of n vertebra instances in an image, spatially ordered from top to bottom. $C = \{C1, C2, \dots, C7, T1 \dots T12, L1, \dots L5\}$ defines all 24 vertebra labels (without T13 and L6) with C_i corresponding to the i -th entry in C . We define the set of possible vertebra transitions label $T = \{\text{None, Last Cervical, First Thoracic, Last Thoracic, First Lumbar, Last Lumbar}\}$. With i corresponding to the i -th vertebra instance V_i , j as the index for the vertebra label C_j , and k as the index for the vertebra transition set T_k , then the classifier outputs normalized vertebra class softmax scores c_{ij} , normalized region prediction scores r_{ij} (where each region output is mapped onto each vertebra class of that region, e.g. the cervical region output is applied to C1 through C7), normalized vertebra transition scores t_{ik} , and a normalized visibility softmax score $s_i \in [0,1]$.

After classifying a sequence of vertebrae, the output is further processed. We add a Gaussian filter on each softmax output per vertebra and then normalize the vector by dividing it by its length. For the vertebra transition output, we normalize over the different classes (i.e., the more vertebrae the model predicts to be the last thoracic one, the lower the values for each of those predictions). Then we combine the outputs of our classification model into a cost function. The intuition is to find the sequence with the maximum agreement between the different classification outputs across the vertebrae.

With w_C, w_R, w_T being weights for the corresponding classifier output types c_{ij}, r_{ij} , and t_{ik} , we calculate the cost for each instance V_i and class C_j with

$$\text{labelcost}(i, j) = (c_{ij} \cdot w_C) + (r_{ij} \cdot w_R)$$

This way, we can define a labelcost matrix $L^{(n \times 24)}$ where $L_{ij} = \text{labelcost}(i, j)$. The vertebra transition cost is utilized differently since it depends on the rest of the path.

Given a candidate path p as a sequence of vertebra class index labels $[j_0, j_1, \dots, j_n]$ with $p_i \in C$ being the i -th entry of p , we define the vertebra transition cost for an individual vertebra V_i as:

$$\text{transcondition}(p, i, k) = t_{ik} \text{ if } T_k \text{ fulfilled in } p; 0, \text{ otherwise}$$

$$\text{transcost}(p, i) = w_T \cdot \sum_{k \in T} \text{transcondition}(p, i, k)$$

The classifier predicts the visibility s_i , which we utilize to weigh each vertebra. The less visible the model predicts the vertebra to be, the less it impacts the sequence predictor. Thus, we can compute the cost of a whole path by using the visibility output and accumulating the individual costs with

$$\text{pathcost}(p) = - \sum_i^n s_i \cdot (\text{transcost}(p, i) + L_i, p_i)$$

Thus, for all $p \in P$ with P being the set of all possible paths, the sequence predictor solves the minimization problem and finds the solution path X by

$$X = \text{argmin}_{p \in P} \text{pathcost}(p)$$

Once we have computed X , all we need to do is replace the class indices p_i with the actual class labels C_{p_i} , then relabel occurrences of two consecutive T12s to [T12, T13] and two consecutive L5s to [L5, L6] to receive the final result.

Since the T2w and CT model solve the same task, we calibrated the sequence predictor weights w_C , w_R , and w_T based on both models on a $\sim 10\%$ split from the training set. Based on this calibration, we set $w_C = 0.9$, $w_R = 1.1$, and $w_T = 0.6$. We observed these to be quite stable regardless of the split utilized.

For training the classifiers, we employ the parameters as shown in Supplemental Table 1.

Appendix B: Ablation studies

Experiments

As additional ablations, we evaluate the robustness on VERIDAH by stratifying the test data. With available annotations regarding stump ribs and scoliosis in the NAKO MRI dataset, and the availability of fracture and stump rib ratings in the CT dataset, we compare the performance of the VERIDAH models on the whole vs. only on anomalous subsets of the test data. Additionally, for MRI, we stratify based on the four most common thoracolumbar enumeration combinations. This was not possible for CT since they have limited FOV and thus, we cannot compute the total enumeration.

Furthermore, we analyze the performance of VERIDAH if we introduce an artificial cost to the paths containing enumeration anomalies similar to Meng et al. (26). The proposed VERIDAH method always predicts a sequence that consists of consecutive vertebrae. In practice, faulty methods to extract the positions of the vertebrae or imaging issues (like artifacts or signal loss) might lead to a missing vertebra in the detection process. Thus, we analyze how VERIDAH would perform in these circumstances. First, we ran the same setup on the test data without adding gaps in the reference data but removing the consecutive constraint on the sequence predictor. This means that the reference data does not contain gaps, but our method can now predict gaps. Then we randomly introduce a gap in each test subject by removing one random vertebra from the reference labels. We rerun the same experiment, showing how well VERIDAH performs in correctly predicting gaps. In practice, the availability of these enumeration anomaly labels is sparse. To showcase the ability of VERIDAH to predict TEA and LEA even though it has never seen those labeled cases during training, we retrain our approach and remove all TEA and LEA annotations by always assuming twelve thoracic vertebrae. This effectively means we relabel all T13 vertebrae to L1 and the L1 of T11 cases to T12 and shift the remaining lumbar vertebra labels accordingly. We evaluate our unmodified test data, thus including the TEA and LEA labels.

Results

Supplemental Table 2 compares the performance of the VERIDAH models on the whole test set with unhealthy subsets of the same test set. While both subjects with scoliosis in MRI or those with at least one fracture in CT do lower the average performance of the VERIDAH models, those results are not statistically significant (PL perc. fractures p-value = 0.18, stump rib p-value = 0.08). For MRI, the cases with a thoracolumbar stump rib reduce the VERIDAH performance significantly (PL perc., p-value < 0.001), while the existence of scoliosis does not (PL perc, p-value = 0.18). Focusing on the metric values, it seems that stump ribs do make it more difficult for both the CT and MRI VERIDAH model (PL perc. from 98.3 to 93.81, and 99.18 to 93.48), respectively. Supplemental Table 3 shows the MRI metrics on the four most common enumeration combinations in the test data. Compared to the whole test set, only the 13/5 combination shows statistical significance (p=0.003). Although having a TEA precision of 100, it seems the VERIDAH model for MRI struggles with these cases the most.

Meng et al. (26) add an artificial penalty to the prediction of TEA and LEA cases, making it less likely to predict these. We investigate this by introducing a cost change of γ for paths containing a TEA or LEA during our sequence predictor. A positive γ makes the model predict more TEA and LEA, while a negative cost introduces more difficulty for the model to predict anomalous sequences. With $\gamma \in [-2, 2]$, we report the changes in metrics using VERIDAH on our test data in Supplemental Figure 3. At $\gamma \leq -1.0$, there is a huge drop in performance. This is probably because at this point, the model only predicts anomalous sequences. Since most subjects have no enumeration anomalies, the sequence predictor has catastrophic performance. With $\gamma \geq 1.25$, the cost of anomalies is so high that it rarely predicts anomalous

sequences, thus leading to lower TEA and LEA recall. Introducing any such cost (regardless of positive or negative) negatively influences the sequence predictor. There is one notable exception in MRI. With $\gamma = -0.25$, there is a slight increase in the TEA recall. from 86.59 to 87.8 while the PL perc. decreases from 98.2 to 97.4. We deem this tradeoff unworthy. We thus refrain from using such a cost in our main approach.

With a punishment cost of ≥ 1.0 , the model never predicts any gaps on the MRI test data. Thus, performance is unchanged. The lower the cost, the worse the performance across all three metrics (see Supplemental Figure 4). This indicates that allowing our approach to predict non-consecutive label sequences lead to predicting gaps where none exist. When we introduce random but fixed gaps into our MRI test data by just removing vertebra locations from the reference and re-running this experiment, we achieve a different result (see Supplemental Figure 5). We observe that VERIDAH achieves optimal results if we keep the skip penalty to zero in the case of actual gaps in the data. The performance drops significantly once we punish the model for predicting gaps in the label sequence, since each subject contains gaps. Interestingly, the TEA rec. has a small peak around the interval $[0.75, 1.00]$ in the MRI sequence, for which we can offer no explanation other than random noise.

All in all, this means that even if subjects have missing vertebrae, if we allow VERIDAH to predict gaps in the sequence, it is correct (including the gaps) in 85.7% of MRI subjects, with a TEA rec. of 69.51% and LEA rec. of 83.45%. For CT, VERIDAH achieves perfect labels in 85.21% of subjects, with a TEA rec. of 88.89% and LEA rec. of 47.22%.

The metrics we achieved with VERIDAH trained on zero correctly annotated TEA and LEA cases are reported in Supplemental Table 4. We again compare using all four classification heads with the sequence predictor (4H-P-NA) against using only the vertebra classification output with the sequence predictor (VH-P-NA). This shows that even if we relabel our reference to count to twelve thoracic vertebrae, VERIDAH still detects about a third of the TEA. This highlights how our method is able to derive important aspects from the different classification outputs despite not having seen any correctly labeled anomalous cases during training.

Appendix C: Hardware and Computation Times

We trained and executed this study with more than 100 GB Ram available, an Nvidia A40, and a AMD EPYC 741 as CPU. To reasonably run VERIDAH, one approximately requires 4 GB of VRAM. With our study hardware setup, one training until convergence took roughly one GPU day on the NVIDIA A40. Inference time with VERIDAH took about 2.5 seconds per image (~24 samples per minute).